DOCUMENT RESUME

ED 250 391 TM 840 742

AUTHOR Milazzo, Patricia A.; Buchanan, Aaron D.
TITLE Equating Instructional Accomplishment Inventories and Standardized Achievement Tests.
INSTITUTION Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO SWRL-TR-77
PUB DATE 12 Apr 82
CONTRACT NEC-00-3-0064
NOTE 30p.
PUB TYPE Reports - Research/Technical (143)

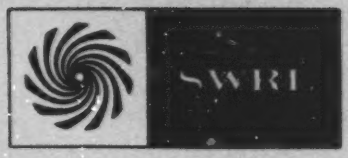
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Criterion Referenced Tests; Elementary Education; *Equated Scores; Grade 3; Grade 6; Research Methodology; *Standardized Tests; Statistical Analysis; Test Items; Test Theory; *Test Validity
IDENTIFIERS Comprehensive Tests of Basic Skills; *Instructional Accomplishment Inventories; Survey of Essential Skills

ABSTRACT

Standardized achievement tests and instructional accomplishment inventories involve different methodologies and cannot be equated by using conventional psychometric methods. Instructional accomplishment inventories are descriptive, and are designed to reflect the scope, sequence, and skills and emphasis in a particular instructional program. Standardized achievement tests are designed to discriminate between students and do not represent actual instruction. These tests can be equated using a qualitative method which requires a matching of instrument structures at three levels: general instrument, subcatagories, and items. The analysis is performed in sequence at each level to show the correspondence between skills reflected in the instrument. An example of qualitative equating for the Comprehensive Tests of Basic Skills and the Los Angeles City Schools' Survey of Essential Skills in reading and mathematics for grades 3 and 6 is given. Qualitative analysis may reveal that there is no meaningful basis for statistical equating. If the testing instruments do have a qualitative relationship, the statistical relationship between the instruments takes on a better informed meaning. (BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED250391



SWRL EDUCATIONAL RESEARCH AND DEVELOPMENT

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. A. Milazzo

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 742

Equating Instructional Accomplishment Inventories and Standardized Achievement Tests

BEST COPY AVAILABLE

This document has been distributed to a limited audience for a limited purpose. It is not published. Copies may be made only with the written permission of SWRL Educational Research and Development, 4665 Lampson Avenue, Los Alamitos, California 90720. The work upon which this document is based was performed pursuant to Contract NE-0-00-3-0064 with the National Institute of Education. SWRL reports do not necessarily reflect the opinions or policies of the sponsors of SWRL R&D.

SWRL EDUCATIONAL RESEARCH AND DEVELOPMENT

TECHNICAL REPORT 77

April 12, 1982

EQUATING INSTRUCTIONAL ACCOMPLISHMENT INVENTORIES AND STANDARDIZED ACHIEVEMENT TESTS

Patricia A. Milazzo and Aaron D. Buchanan

ABSTRACT

A method for qualitatively equating achievement measures is described and illustrated by an example applied to elementary reading and mathematics. The method complements and extends conventional methodology for quantitatively equating educational testing instruments.

EQUATING INSTRUCTIONAL ACCOMPLISHMENT INVENTORIES AND STANDARDIZED ACHIEVEMENT TESTS

Patricia A. Milazzo and Aaron D. Buchanan

Statistical procedures derived from classical psychometric theory and practice underlie a methodology that has been used for many years to equate all variants of standardized achievement tests (see Thorndike, 1971). The methodology also has been applied successfully to criterion-referenced achievement tests, with appropriate caveats about the distinction between "equivalence" and "comparability." Instructional accomplishment inventories (e.g., SWRL's Proficiency Verification Systems, Los Angeles' Survey of Essential Skills, Sacramento's Proficiency Survey, the District of Columbia's Competency Based Assessment), however, are fundamentally different from both norm-referenced and criterion-referenced tests in ways that make the conventional statistical equating information inadequate in the absence of analytic equating information.

The present paper presumes the reader is familiar with standard statistical equating and with the general methodology underlying standardized achievement tests. The paper describes briefly the general methodology underlying instructional accomplishment inventories. It then outlines the relationship of both standardized achievement tests and instructional accomplishment inventories to instruction. With this information as background, the paper describes a method for performing qualitative equating and illustrates the method with a sample analysis.

Methodology Underlying Instructional Accomplishment Inventories

The determining difference between standardized achievement tests and instructional accomplishment inventories lies in the distinction between psychometric methodology (applicable to standardized achievement tests) and survey research methodology (applicable to instructional accomplishment inventories). Although some pertinent principles and features of psychometric technology may be applied in the development

and use of instructional accomplishment inventories, many key aspects of psychometric methodology (e.g., validity, reliability, item analysis statistics) are not directly applicable. In place of psychometric methodology, instructional accomplishment inventories make use of methodology derived from social science survey research. The critical features of instructional accomplishment inventories are representation of instructional scope and sequence, representativeness of performance modes that are directly familiar to the respondent, and clarity of question and response. These factors derive from two logical tenets fundamental to survey research: ask questions that are most representative of an area of interest, and remove as much ambiguity as possible from every question and every set of responses (Babbie, 1976; Goode & Hatt, 1952). An example from survey research helps to explain the survey approach. If one is interested in knowing how the 1984 Republican presential primary is shaping up, a survey item could look like this:

Who would you vote for in the 1984 Republican presidential primary?

a. ____ Ronald Reagan
b. ____ George Bush
c. ____ Howard Baker
d. ____ Don't know

For the sake of the example, pretend that 80 percent of the respondents choose a, 15 percent choose b, 5 percent c, and 0 percent d. From a survey perspective, the item would be reviewed on the following basis: Are the question and the responses unambiguous? Is the substance of both the question and the responses relevant to the area of interest, the Republican presidential primary? Survey items are frequently refined if questions like these are answered with "no." However, there is no need for alarm, at least not from a measurement point of view, if most respondents load on the "a" response. The item is intended to obtain descriptive information at a specific point in time, and readers may, or may not, decide to take some campaign action on the basis of this information.

From a psychometric perspective, if this survey question were treated like a test item, the question and/or responses would most likely be adjusted in order to relieve the loading on the "a" response. A number of standard techniques are possible for distributing response choices. For example, an "e" response for another likely candidate could be added; some ambiguity could be added by using names such as Donald Regan, by adding a fictitious name such as Ronald Bush, or a plausible but irrelevant name such as Ted Kennedy. These kinds of adjustments should move response choices around somewhat, taking the load off of response "a." By making this type of psychometric adjustment, items become better "test" items, but the descriptive power of the information is seriously reduced.

Instructional accomplishment inventories have a similar power to describe performance on specific skills at specific points in the schooling year. The overriding concern is to reflect accurately the scope, sequence and emphasis of skills represented directly in instruction and practice. Information gathered from surveys has only incidental utility for making discriminations among respondents; that is not their purpose. The fundamental purpose of an instructional accomplishment inventory is to describe, not to discriminate.

Given this distinction, the equating of instructional accomplishment inventories and standardized achievement tests might seem an unnecessary exercise. However, the matter cannot be dismissed so easily. Standardized achievement tests have come to be the standard by which the profession and the public insist that instructional programs, and eventually schools, be evaluated. For this reason, any alternative is obligated to justify itself against this standard. (E.g., "These results are all well and good, but how would the kids do on a standardized test?") Until the considerations involved in equating the two types of instruments are understood, standardized achievement tests will continue to provide the exclusive gauge of instructional program effectiveness, despite their acknowledged deficiency for this purpose (Buros, 1977; Tyler, 1971; Nader, 1979; Airasian, 1979).

Relationship of Standardized Achievement Tests to Instruction

Whereas instruction on a given skill is designed to close the gaps between students in what they learn, standardized tests are designed to identify gaps between what students have learned, to spread students out relative to each other:

> ". . . It would therefore be a mistake to conclude that an item with a relatively small percent of pupils, say 35 percent, answering correctly represents a skill which needs immediate attention. *This could be an item which represents a level of performance that few pupils should be expected to master* (italics added)." (Houghton Mifflin Company, Boston: Iowa Test of Basic Skills, Item Performance Analysis, Forms 5 and 6, Grade 6, Level 12, 9-67535, Copyright 1971.)

When instruction has been highly successful in teaching a skill to nearly all students, which is often true in the elementary grades, a standardized test cannot align well with the skill because performance scores would be "too high" to discriminate between students. To obtain the intended discrimination, the test is made more difficult than instruction would merit. By adding a level of difficulty (sometimes two or three levels of difficulty) to test items, test scores can be made to spread out in a downward direction. Hence the tendency for these tests to underestimate the instructional accomplishments of students with the least instructional opportunity or success. Similarly, when instruction is largely unsuccessful in teaching a skill to nearly all students, the standardized test cannot align well with the skill that a large majority of students have not learned, because performance scores will tend to cluster in the low ranges. In this case, widespread instructional mal-achievement makes it difficult to discriminate between students. A level or two of difficulty can be removed from test items, forcing scores to spread out in an upward direction. Hence the tendency for standardized tests to overestimate the instructional accomplishments of students with the most instructional opportunity or success. This practice occurs often in grades four, five and six, where the scope and substance of instruction is frequently very difficult.

Standardized tests align best with instruction that is partially successful, that teaches specific skills to just some, but not all, students. When instruction itself tends to spread performance scores out from high to low on specific skills, there is no need to tinker with item difficulty, since it is possible to discriminate between students by aligning with instruction on these skills. However, data on what is taught and learned indicate that there are a limited number of skills at each grade level which actually fit the paradigm of partial learning; i.e., are learned well by some students, learned partly by other students, and not learned by still other students. (See Los Angeles Unified School District, 1979; Sacramento City Unified School District, 1979; Buchanan & Milazzo, 1978.) Moreover, a close look at such skills, which do show differential performance scores across groups, tends to mitigate a good bit of alarm about low performance scores. Many of these skills are indirect extensions of skills learned through direct instruction and practice, and performance requires transferring knowledge about a learned skill to a new application; the skills may be embedded in rare, or at least unusual, performance formats; or the skills may involve a high degree of ambiguity about what is being taught and what performances are expected . . .. But, grade-by-grade, the skills do not often represent a large investment of instructional time, or a high expectation for mastery. Other skills are ones that are taught across several grades and, sooner or later, they are learned by most students along the way. The lowest scores on these skills occur at the earliest grade levels, where the instructional investment is low and the intention is to introduce skills which will be thoroughly taught at the next higher grade. Students who do best in learning such skills are not absent much; they pay attention in class; they do independent seat work and homework; they are troubled less by problems outside the classroom; they are consistently high achievers; all characteristics which are not very surprising.

On the other hand, when one looks at skills that have the largest commitment of lesson space at each grade and the greatest impact on

grade-by-grade achievement, the effects of instruction tend to be much more common across all groups of students, including students who are considered to be "low achievers."

Relationship of Instructional Accomplishment Inventories to Instruction

The model that underlies instructional accomplishment inventories has no pre-established requirement for item difficulty or for the shape of a performance score distribution. Because the model is rigorously descriptive, the critical factor for survey instruments is goodness of fit with the scope, sequence, and emphasis of skills taught in a particular program of instruction. The structure and substance of an instructional accomplishment inventory are formed and justified on the basis of that fit, independent of the statistical characteristics of items or scores.

The scope, substance, and format of instructional accomplishment inventories are derived from the instructional objectives and resources to which a district is committed. Item formats are then designed to reflect, as much as possible, highly familiar, representative practice formats from that instruction. Weight given to the various subcategories in each major skill category of each subject area is determined by the lesson emphasis that the skills receive in a district's program of instruction.

Unlike standardized achievement tests, goodness of fit for instructional accomplishment inventories does not depend on how successful instruction has been. If substantial amounts of lesson space are dedicated to teaching specific skills, then the skills should be represented in an inventory, regardless of how this practice affects the overall performance score distribution.

Description and Illustration of a Method for Qualitative Equating

The method requires a matching of the structure of the instruments to be equated at three levels: general instrument, subcategories, and

items. The analysis is performed in sequence at each level to show the correspondence between the skills reflected in the instruments.

The method is most conveniently explained via an example derived from SWRL's collaborative effort with Los Angeles City Schools to help the District implement its grade-by-grade promotion policy for elementary school students. In this connection, the District administers annually an instruction-based accomplishment inventory (Survey of Essential Skills-SES) to more than 300,000 elementary students. The District also wanted to use the SES as a part of the evaluation of its ESEA Title I program. Federal regulations permitted the District to use the SES for Title I reporting purposes, if the SES were equated with a standardized achievement test.

The example presented here illustrates the qualitative equating that was done for the California Test of Basic Skills (CTBS) and the SES in the subject areas of reading and mathematics for grades 3 and 6.

Stage One: Equating General Categories

The first operation in this stage is strictly descriptive. It provides a simple listing of the querying and reporting categories that are named in each instrument Tables 1 and 2 show the structures of CTBS and SES for grades 3 and 6 respectively. For example, the CTBS instrument that is recommended for use at grade 3 shows two general querying and reporting categories in reading and two in mathematics. The SES instrument for grade 3 shows five such categories for reading and eight for mathematics.

In the second operation in this stage the broad skill categories from each instrument are compared and "matched." The task is obviously easiest to accomplish where the two instruments have skills categories with identical titles. For example, in Tables 1 and 2, CTBS has a broad skill category labeled vocabulary, so does the SES; CTBS has a broad skill category labeled comprehension, so does the SES. Therefore, at

the grossest level of comparison, CTBS and SES reading instruments for these grades show at least two broad skill categories that are nominally matched.

Occasionally, broad skill categories may not match exactly in name, but the constructs are alike. For example, if one instrument has the skill category "word meaning," and the second instrument has the skill category "vocabulary," these constructs would be considered a match. Similarly, the broad skill categories on one instrument may encompass several of the broad skill categories on a second instrument. This occurs in the mathematics portions of Tables 1 and 2. Although there may be no nominal counterpart from one instrument to the next, if constructs are similar, "matching" can be accomplished by simply collapsing or unfolding the broad skill categories on one of the instruments. For example, CTBS has a reporting category labeled "computation," for which there is no direct counterpart on the SES. However, the SES does have two querying and reporting categories that are clearly computation, "Addition and Subtraction of Whole Numbers," and "Multiplication and Division Facts." Without creating new constructs, the two instruments can be linked at this level by collapsing the two SES computation categories, or unfolding the broader CTBS category.

The results of this first stage of analysis will reveal the general relationships between the instruments, according to "structure" (the number and allocation of items) and to "substance" (types of broad skill categories assessed and reported). The intention in this earliest stage of comparison should be to apply a liberal metric for equating, to accept as much of the total instruments as possible for the next level of the analysis.

Stage Two: Subcategories Within General Skills Categories Which Are Matched

The second stage of analysis focuses on those broad skills categories which were found to be nominally alike in stage one. The intention is to

Table 1
Comparison of General Skill Areas
on CTBS-Level 1, Form S and SES Grade Three

Skill Areas	CTBS	SES
Reading	85 Items	53 Items
Vocabulary	X	X
Comprehension	X	X
Decoding		X
Structural analysis		X
Location/study skills		X
Mathematics	98 Items	60 Items
Addition and subtraction of whole numbers	X	X
Multiplication and division facts	X (reported as computation)	X
Numeration		X
Fractional numbers	X	X
Geometry	X (reported as concepts and applications)	X
Measurement	X (reported as concepts and applications)	X
Relations/functions/ statistics	X (reported as concepts and applications)	X
Applications	X (reported as concepts and applications)	X
Total number of general Skill Areas:	5*	13

*Concepts and applications are reported as a single skill area on CTBS.

Table 2
Comparison of General Skill Areas
on CTBS Level 2, Form S and SES Grade Six

Skill Areas	CTBS	SES
Reading	85 Items	62 Items
Vocabulary	X	X
Comprehension	X	X
Decoding		X
Structural analysis		X
Reference/study skills		X
Mathematics	98 Items	63 Items
Addition and subtraction of whole numbers	X	X
Multiplication and division of whole numbers	X (reported as computation)	X
Computation with fractions and decimals		X
Numeration		X
Fractional numbers	X	X
Geometry	(reported as	X
Measurement	concepts and	X
Relations/functions/statistics	applications)	X
Applications	X	X
Total number of general Skill Areas:	5*	14

*Concepts and applications are reported as a single skill area on CTBS.

achieve a finer level of detail than the gross reporting categories, and to begin to integrate the structures of the two instruments.

For convenient comparison and matching, the large number of items in each skills category is organized into more homogeneous subcategories, which would be meaningful across the two instruments at the grade level of concern. At this stage, the primary concern is simply to remove a layer of structural complexity represented by the two instruments, prior to establishing linkages between individual items. In our example, it was possible to use existing indexes in reading or mathematics which have enough surface detail to permit a breakdown of large skill constructs, such as vocabulary or computation, into more homogeneous and descriptive subconstructs. Table 3 shows the original subcategorization schema selected for the reading analysis, which was adapted from a much more detailed index (see Fiege-Kollmann, 1977). The single asterisks in Table 3 indicate the subcategories which became meaningful in the actual analysis of CTBS and SES reading instruments at grades three and six (i.e., the subcategories which were actually used by the coders). Table 4 shows the same schema for mathematics, also adapted from a much larger index (see Buchanan, 1976). Occasionally, constructs turn up on reading instruments which are <u>not</u> part of the original schema for a particular broad skill category, such as comprehension. If the instruments being equated are to be described in much detail, these constructs require the addition of separate subcategories to the schema. For this reason, any coding schema that is used should be treated as an open-ended framework, a tool that can be refined as the need arises in the analysis procedure. The double asterisks in Table 3 are an example of subcategories that were added by coders. The literary constructs with double asterisks rarely appear under "comprehension skills" in most conventional indexes. However, they did appear in the CTBS test under comprehension, and they were included in the coding schema.

One coder with background experience in reading instruction, and one with background experience in mathematics instruction, was asked to code

Table 3: Subcategories for Use in Reading Instrument Analysis

VOCABULARY SUBCATEGORIES

* Synonymns, given minimal or no context
* Antonymns
* Definitions
* Words contextually cued (common, homonymns, homographs, multiple meaning)
* Function words (prepositions, pronouns, proforms)
* Figurative language

COMPREHENSION SUBCATEGORIES

* Facts, details
* Sequence of events
* Cause-effect
* Main idea (topic, title)
* Conclusions
* Predictions/judgments
  Following directions
  Comparisons
  Contrasts
  Analogies
* Classification
* Relevant versus irrelevant
** Figurative context/devices
** Quotations
** Poems/poetry elements

* Categories actually used by coders in the analysis

** Categories added by coders and actually used in the analysis

Table 4: Subcategories for Use in Mathematics Instrument Analysis[1]

COMPUTATION

Addition and subtraction of basic facts
Addition and subtraction of 2-digit numbers, no regrouping
* Addition and subtraction of 2-digit numbers, regrouping
* Addition and subtraction of 3-digit numbers, no regrouping
* Addition and subtraction of 3-digit numbers, regrouping
* Addition and subtraction of large numbers
* Multiplication and division facts
* Multiplication by 1-digit multipliers, no regrouping
* Multiplication by 1-digit multipliers, regrouping as necessary
* Multiplication, large numbers, by 2-3-digit multipliers
* Division by 1-2 digit divisors, no regrouping
* Division by multiple of ten
* Division by 1-digit divisors, regrouping as necessary
* Division, large numbers, by 2-3-digit divisors
Addition and subtraction of fractions, like denominators, no regrouping
* Addition and subtraction of fractions, like denominators, regrouping as necessary
Addition and subtraction of fractions, unlike denominators, no regrouping
* Addition and subtraction of fractions, unlike denominators, regrouping as necessary
* Multiplication and division of fractions
* Addition and subtraction of decimals
* Multiplication and division of decimals by whole numbers, 10, 100
Multiplication of decimals by decimals
* Division of decimals by decimals

* Actually used in the analysis

[1]This paper discusses only the computation sections of the two instruments; and, therefore, only the computation portion of the mathematics index is provided here. See Buchanan, 1977, for the complete index.

grades three and six, using the specific subcategories shown in Tables 3 and 4. Coders were told to be consistent in classifying items at both grade levels of both instruments. The critical practice in this stage of the analysis is to apply the categorization schema systematically to both instruments. The procedure is most stable when a coder is responsible for both instruments in one grade level.[2]

Stage Three: Items Within Subcategories Which Are Matched

The third, most specific level of analysis focuses on items in those subcategories which are nominally alike. In the example analysis, two item features were identified as plausible, but not necessarily the only, indicators of item equatability: the skill assessed and the item difficulty (in percent correct). In dealing first with skills actually assessed, it seemed reasonable to assume that items in the same subcategory attended to, more or less, the same skill constructs. Item difficulty, on the other hand, required some additional considerations. The initial task was to establish a range of item difficulties in each subcategory. This was done by arraying difficulty values for each item in the matched subcategories from most difficult to least difficult. Table 5 gives an example of the array for reading comprehension at grade three. Using this type of table, items that fall strictly within overlapping difficulty ranges are analyzed first; then items that are adjacent to this range of difficulty values are analyzed. It is usually reasonable to stay within a range of plus or minus .25 from the extreme values on the strictly overlapping items since this scope is broad enough to permit a qualitative analysis of many items. A reading specialist reviewed and rated the items in both instruments as "more or less similar," or "more or less dissimilar." A mathematics specialist completed the same activity for the mathematics instruments. Items were reviewed on a number of features, such as language

[2]At this point, a methodological note is in order. Researchers often forget that their categorization schemas are constructed, not devined. While some schemas may have more, or less, of a descriptive relationship to instructional materials than others, one would be hard-pressed to identify the "best," or, even more optimistically, the "right" set of constructs. Whatever set is used, methodologically, the fundamental concern should be with systematic application across the two instruments being equated.

Table 5: Array of Item Difficulty in Percent of Right Response for CTBS and SES Reading Instruments at Grade Three

	CTBS		SES			
	Item No.	Ordered Difficulty	Item No.	Ordered Difficulty		
Comprehension Details	11	.32				
	39	.35				
	9	.41				
	12	.55				range +/- .25
	18	.57				
	38	.57				
	37	.59				
	36	.61				
	26	.62				
	17	.63				
	10	.64				
	1	.65				
	29	.65				
	21	.66				
	28	.67				
	16	.71				
			35	.72	▲ overlapping difficulty ranges	
			40	.73		
	5	.74				
			33	.79		
Comprehension Sequence	40	.40				
			37	.58	▲ overlapping difficulty ranges	range +/- .25
	21	.61				
	8	.63				
			36	.66		
			38	.68		
	6	.75				
	2	.78				
Comprehension Conclusions	32	.46				
	33	.48				
	19	.51				
	24	.51				
	25	.51				
	13	.53				
	23	.53				range +/- .25
	22	.54				
	30	.58				
	31	.62				
	35	.64				
	4	.71				
			41	.75	▲	
			38	.77		
	7	.80				

level, syntax, semantics, format, contextual clarity, response discriminations, specificity of the subskill (e.g., regrouping with zeros), and so on. The intention was to be liberal in identifying items which might be thought similar, thus producing as large a pool of "linked" items as possible.

Interpretation of Qualitative Equating Information

In the sample analysis, the similarities between standardized achievement tests and instructional accomplishment inventories decrease as the skills, and eventually the items, become more specific. At the most general analytical level, that is on the surface, the two instruments look somewhat alike. This first cut comparison, however, reveals some interesting differences about the nature of the subject matters. For example, reading is a subject matter where most of the specific technical "reading" skills are pretty much taught in grades one, two, and early grade three. After that time, students do not learn to read, as much as they "read to learn." That is, they apply their reading skills in order to read and understand longer, more sophisticated texts. Hence there is a small number of broad reading constructs represented in both CTBS and SES at grades three and six, and the general constructs can be maintained grade-by-grade. The content of mathematics instruction is different. Through grade six, there are specific, technical mathematics skills that students are expected to learn. These kinds of technical skills actually increase in number in grades four and five, while applications tend to have a low profile throughout the intermediate grades in most programs of instruction. Hence, there are many mathematics constructs represented in both CTBS and SES at grades three and six. These general constructs do more changing grade-by-grade in mathematics than in reading, because there is less emphasis on process and more emphasis on specific, technical skills.

At the general instrument level of analysis, the querying categories of both instruments, although not the reporting categories, seem to focus on about the same general skills. In a standardized achievement test,

many skills tend to be collapsed into one skill category for reporting; whereas in an instructional accomplishment inventory, more homogeneous skills categories are reported separately whenever possible.

Tables 1 and 2 give qualitative indicators that the statistical equating of the SES and CTBS for reading has a tenuous qualitative basis at best. More than half of the broad skill categories which are assessed and reported in the SES are not represented in the CTBS reading test, unlike the mathematics test where all of the broad skill categories in the SES are collapsed under even broader CTBS categories. The SES in reading represents large chunks of instruction which simply are not reflected in CTBS reading. This means that more than half of the SES and CTBS items are not equatable in any way that reflects actual skills assessed. Specifically, 30 of the 53 SES reading items in grade three, and 31 of the 62 SES reading items in grade six, fall in general querying and reporting categories that have no counterpart in the CTBS reading test. The remaining appearance of equatability at the general instrument level is an illusion created by categories with the same names, but substantively different representation. The illusion became clear as the analysis moved to finer levels of specificity.

At the subcategory level of analysis, the fit between CTBS and SES reading instruments nearly disappeared, and the fit between the mathematics instruments began to strain. Tables 6 and 7 compare the reading instruments at grades three and six, respectively. In the vocabulary test, the the tables show that CTBS assesses only one subskill (coded as synonyms), using a very large number of items with an identical format (test length being an important factor in obtaining maximum subtest reliability). SES, on the other hand, represents the broad scope and emphasis of vocabulary instruction at these grade levels by surveying several subcategories of vocabulary skills. This procedure provides just two synonym items in SES that might possibly be equated to the 40 synonym items in CTBS. A total of ten equatable reading items were identified at grade three; and a total of seven equatable items were identified at grade six. There are no more items.

Table 6: Comparison of CTBS and SES Subcategories in Vocabulary and Comprehension Skill Categories for Grade 3

Categories	CTBS	SES
Vocabulary	**40 items**	**10 items**
Prepositions		X
Synonyms	X	X (2 items)
Context clues		X
Definitions		X
Antonyms		X
Comprehension	**45 items**	**13 items**
Details	X	X (3 items)
Sequence	X	X (3 items)
Drawing conclusions	X	X (2 items)
Main idea	X	
Classification		X
Total number of subcategories:	5	9
Total number of SES items in similar subcategories:		(10)

Table 7: Comparison of CTBS and SES Subcategories in Vocabulary and Comprehension Skill Categories for Grade Six

Categories	CTBS	SES
Vocabulary	**40 items**	**12 items**
Definitions		X
Synonyms	X	X (2 items)
Figurative language		X
Antonyms		X
Comprehension	**45 items**	**19 items**
Drawing conclusions	X	X (1 item)
Prediction		X
Irrelevancy		X
Classification		X
Main idea	X	X (1 item)
Details	X	X (2 items)
Cause/effect		X
Sequence	X	X (1 item)
Quotations		X
Poetry elements	X	
Figurative context	X	
Total number of subcategories:	7	13
Total number of SES items in similar subcategories:		(7)

There is little reason to carry the analysis of the reading instruments any further. Even if all ten items at grade three and all seven items at grade six identified as equatable at the subcategory level were to prove equatable at the item level of qualitative analysis, it makes no sense to equate statistically seven or ten items of a 50 or 60 item instrument to 85 items on another instrument. In fact, not all ten SES reading items at grade three, nor all seven at grade six, have a corresponding item form in CTBS reading. For example, neither of the SES synonym items at these grade levels are qualitatively "similar" to CTBS items, not even where item difficulty values are similar. All 40 of the CTBS synonym items have the same performance format: A two- to four-word ambiguous phrase is the "stimulus," with a target word that is frequently above grade level, and with four response choices, including two or three acceptable, if not "best," answers. SES synonym items are set intentionally in disambiguating two- to four-sentence contexts, with three or four response choices, and much more emphasis on "right" answers. Even using a liberal metric for equating items, these kinds of performance formats are not similar. Most often, performance scores won't be similar either.

CTBS--Level 1, Form S	SES--Grade Three
20. good idea ○ example ○ fact ○ mood ○ thought P=.53	24. John wants to buy that coat. He does not care about the price. Which word means the *same* as the underlined word? market ○ order ○ cost ○ P=.76

The qualitative equating story for mathematics in the present study has about the same ending as that for reading. It just takes longer to tell, and the final break is not obvious until the analysis reaches the item level. However, there is an interesting twist in the tale when it

is told for mathematics, and so an abbreviated version of that story is provided here. For this purpose, the analysis can focus on the broad area of computation. Table 8 shows the subcategory breakdown for the grade three instrument.

The asterisks in Table 8 point out the interesting nature of standardized testing. In Table 8, the asterisks indicate that four of the nine subcategories assessed in CTBS for grade three represent skills that are well beyond mainline grade three instruction. From a psychometric perspective, this procedure makes sense. Instruction throughout the primary grades tends to be quite successful, i.e., many more students learn more of the skills that are taught about on schedule than they do in later grades. Therefore, performance scores on the mainline skills will have some tendency to cluster in the middle to high score ranges at grades one, two, and three, and a level of difficulty has to be added to the CTBS test to distribute scores in a downward direction. In reading, this is frequently accomplished by manipulating the language (e.g., a complex syntax may be used, vocabulary that is above grade level may be included in items, etc.). With computation, it is difficult to affect scores by manipulating the language, since most computation items are basically language free. What can be done to improve the discrimination power of the test is to incorporate a large number of items on skills that may be only introduced at the grade level but are taught and learned at a higher grade level. These kinds of skills occur frequently in mathematics instruction because of the linear characteristics of that instruction. For example, students are taught to add and subtract with regrouping on small numbers before they are taught to add and subtract with regrouping on large numbers. The former skill is expected to come on line by the end of grade three and the latter skill by the end of grade four. At the same time, most comprehensive programs for mathematics instruction will include a small number of lessons near the end of grade three to briefly introduce the skill.

Table 8

Comparison of CTBS and SES Subcategories
in the Computation Skills Category
for Grade Three

Categories	CTBS	SES
Computation	48 items	14 items
Addition and subtraction of numbers to 2-digits, regrouping as necessary	X	X (6 items)
Addition and subtraction of numbers to 3-digits, no regrouping	X	X (2 items)
*Addition and subtraction of numbers to 3-digits, regrouping as necessary	X	
*Addition and subtraction large numbers	X	
Multiplication and division facts	X	X (6 items)
Multiplication by 1-digit multipliers, no regrouping	X	
Multiplication by 1-digit multipliers, regrouping as necessary	X	
**Division by 1-digit divisors, no regrouping	X	
**Division by multiple of 10	X	
Total number of subcategories in this skills category:	9	3
Total number of SES items in similar subcategories:		(14)

*Introduced late in grade three, retaught seriously in grade four.

The fit between standardized achievement tests and instructional accomplishment inventories seems to improve for mathematics in the intermediate grades, as illustrated in Table 9. Finally, at least for this large subtest area, there seems to be a fairly large number of equatable CTBS and SES items. This makes sense from both a psychometric and an instructional perspective. By grade six, the skills involved in mathematics instruction have become much more difficult for most students. Therefore, items taken directly from instruction and practice have a natural tendency to have moderate difficulty levels, which is necessary from a psychometric perspective, so there is no need to "add" any difficulty. To the contrary, since most students are going to cluster in the middle-to-low score ranges on skills that are taught in instruction, a level of difficulty often has to be removed from the computation subtest in order to spread scores out in an upward direction. This condition makes for an interesting twist in the standardized test for mathematics in the intermediate grades. A large number of items are incorporated that represent skills which are mainly taught and learned at lower grade levels. The asterisks in Table 9 indicate a number of subcategories in the CTBS test intended for use in grade six which represent instructional content that is somewhat below mainline grade six computation instruction. Thus, in the intermediate grades, the poor fit in computation subcategories is due largely to CTBS items that are below grade level, a very different condition from grade three.

This "twist" identifies just one area of poor fit between CTBS and SES mathematics instruments, and it is mostly confined to computation skills. In the other skill areas, statistical equating becomes as suspect as it does for reading because of the number of SES items that are lost before and after the comparison process begins. By the time an item analysis is extended to the entire 98-items in the CTBS test and the 60 or so items in the SES instrument, the "pool" of equatable items is very small in proportion to total test length in both grades three and six.

Table 9

Comparison of CTBS and SES Subcategories
in the Computation Skills Category
for Grade Six

Categories	CTBS	SES
Computation	48 items	21 items
*Addition and subtraction of numbers to 3-digits, regrouping	X	
Addition and subtraction of large numbers	X	X (5 items)
*Multiplication and division facts	X	
*Multiplication by 1-digit multipliers, regrouping as necessary	X	
Multiplication, large numbers by 2-3 digit multipliers	X	X (2 items)
Division by 1-digit divisors, regrouping as necessary	X	X (2 items)
Division, large numbers by 2-3 digit divisors	X	X (1 item)
Addition and subtraction of fractions, like denominators, regrouping as necessary	X	X (3 items)
Addition of fractions, unlike denominators, regrouping as necessary	X	
Multiplication and division of fractions	X	X (2 items)
Addition and subtraction of decimals	X	X (2 items)
Multiplication and division of decimals by whole number, 10, 100	X	X (4 items)
Division of decimals by decimals	X	
Total number of subcategories in this general skills category:	13	6
Total number of SES items in similar subcategories:		(21)

*Reviewed early in grade six, but taught seriously in grades three, four, and five.

For mathematics, unlike reading, much of the incompatibility between SES and CTBS becomes most apparent at the item level. For example, in the primary grades, CTBS items, unlike SES items, tend to focus on the moderate to difficult nuances of skills (e.g., regrouping with zeros, or regrouping in two places, etc.). The simple nuances, which often receive the heaviest emphasis in instruction, tend to be represented with a few items in CTBS. A typical example of item incompatibility is shown below, and it demonstrates this focus on different nuances of a skill.

CTBS--Level 1, Form S	SES--Grade 3
Mr. Smith washed his car. The two clocks show you when he started and when he finished. At what time did he finish? START FINISH ○ 6:40 ○ 7:30 ○ 8:00 ○ 8:30 P = .60	Mark the time. ○ 6:10 ○ 10:30 ○ 10:06 P = .80

These measurement items belong nominally to the same subcategory--time. But the items, as well as the performance scores (P values), are clearly not comparable.

General Applicability of Qualitative Equating

This paper has presented a method for qualitative equating, a matter which has not been studied seriously until now. The presentation has focused on instructional accomplishment inventories and standardized achievement tests. However, the method has wider applicability to all varieties of instruments. Traditionally, test developers have described

the statistical relationships between instruments, but they almost never describe the meaningfulness of the relationship. By preceding quantitative analysis with a qualitative analysis, researchers can provide a logical foundation for equating two instruments. In some instances, the results of the qualitative analysis will reveal that there is no meaningful basis for conducting statistical equating, eliminating the necessity for a statistical analysis. In other applications, the qualitative results will support a quantitative equating operation.

In any case, the method presented should refine the paradigm for quantitatively equating testing instruments. Researchers now have another question to ask before proceeding with statistical operations: "How extensive is the qualitative basis for equating the specific instruments?" It is unscientific to assume that the answer will always support quantitative procedures for equating instruments. On the other hand, where it is shown that testing instruments do have a qualitative relationship, the statistical relationship between the instruments takes on a better informed meaning.

References

Airasian, Peter W. "A Perspective on the Uses and Misuses of Standardized Achievement Tests," NCME Measurement in Education, Volume 10, No. 3, Fall, 1979.

Babbie, Earl R. Survey Research Methods, Wadsworth Publishing Company, Belmont, California, 1973.

Buchanan, Aaron D. "Proficiency Verification Systems (PVS): Index of Mathematical Skills," Southwest Regional Laboratory, Technical Note, TN 3-76-01, February 27, 1976.

Buchanan, Aaron D., and Milazzo, Patricia A. "General Pattern in K-6 Mathematics Instruction and Achievement, Report 1: Overall Trends," Southwest Regional Laboratory, Technical Note, TN 3-78-06, March 15, 1978.

Buros, Oscar K. "Fifty years in testing: some reminiscences, criticisms and suggestions," Education Researcher, Volume 6, No. 7, July-August, 1977.

Fiege-Kollmann, Laila. "Proficiency Verification Systems: Index of Reading Skills," Southwest Regional Laboratory, Technical Note, TN 3-77-01, April 1, 1977.

Goode, William J., and Hatt, Paul K. Methods in Social Research, McGraw-Hill Book Company, Inc., 1952.

Nader, Ralph W. Address delivered at the Conference on Testing Reform: Strategies for the 1980's, sponsored by the National Education Association, Washington, DC, January, 1980.

Thorndike, Robert L. Educational Measurement, second edition, American Council on Education, Washington, DC, 1971.

Tyler, Ralph W. "Testing and Achievement Programs--National," The Encyclopedia of Education, Volume 9, pp. 175-170, McMillan Corporation and the Free Press, 1971.

---- "Feasibility Study for Grade-by-Grade Advancement Program," Los Angeles City Unified School District, March, 1979.

---- "Grade Five Proficiency Survey, Feasibility Study," Sacramento City Unified School District.